Research Report

Small-Sample DIF Estimation Using Log-Linear Smoothing: A SIBTEST Application

Gautam Puhan

Tim P. Moses

Lei Yu

Neil J. Dorans

Research & Development

March 2007
RR-07-10

Small-Sample DIF Estimation Using Log-Linear Smoothing: A SIBTEST Application

Gautam Puhan, Tim P. Moses, Lei Yu, and Neil J. Dorans
ETS, Princeton, NJ

March 2007

As part of its educational and social mission and in fulfilling the organization's nonprofit charter and bylaws, ETS has and continues to learn from and also to lead research that furthers educational and measurement research to advance quality and equity in education and assessment for all users of the organization's products and services.

ETS Research Reports provide preliminary and limited dissemination of ETS research prior to publication. To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Abstract

The purpose of the current study was to examine whether log-linear smoothing of observed score distributions in small samples results in more accurate differential item functioning (DIF) estimates under the simultaneous item bias test (SIBTEST) framework. Data from a teacher certification test were analyzed using White candidates in the reference group and African American candidates in the focal group. Smoothed and raw DIF estimates from 100 replications under seven different sample-size conditions were compared to a criterion to determine the effect of smoothing on small-sample DIF estimation. Root-mean-squared deviation and bias were used to evaluate the accuracy of DIF detection in the smoothed versus raw data conditions. Results indicate that, for most studied items, smoothing the raw score distributions reduced variability and bias of the DIF estimates especially in the small-sample-size conditions. Implications of these results for actual testing programs and future directions for research are discussed.

Key words: DIF, small sample, smoothing, subgroups, log-linear smoothing

Acknowledgments

The authors would like to thank Jinming Zhang, Xueli Xu, Dan Eignor, and Sandip Sinharay for their helpful comments and suggestions on an earlier draft of the paper and Kim Fryer for editorial assistance.

Table of Contents

List of Tables

Page

List of Figures

Page

Introduction

Testing programs frequently encounter the problem of dealing with less than optimal sample sizes for conducting statistical operations. This problem is compounded when analyses based on subgroups are conducted, which requires existing small samples to be divided into even smaller samples based on variables such as gender and ethnicity. Differential item functioning (DIF) analyses (Holland & Thayer, 1988) are often used to compare the performance of subgroups on different test items. DIF methods match examinees on total test scores to see if comparable examinees from different subgroups perform the same on individual items. The main idea is that, if members of two groups are comparable in the ability that is measured by the test, then their performance on each item in a test is expected to be similar except for sampling errors. However, if group differences in performance beyond sampling errors are noted, then the item is labeled as performing differentially, and more intensive investigations are carried out to identify the source of these differences (Clauser & Mazor, 1998).

Like other statistical operations, DIF estimates are subject to two general sources of error. *Systematic error* may result from factors such as using a particular statistical method for estimating DIF when another method was clearly more appropriate. For example, using a DIF method that is bassed on item response theory (IRT) when the IRT model does not fit the data may lead to systematic error in the DIF estimation. *Random error* is present whenever a sample from a population of examinees is used to estimate DIF. When the sample is large and representative of the population, then sampling error is small and the DIF estimate is likely to be a close estimate of the population DIF. But when the sample is small and does not reflect the population, the DIF estimate may be very inaccurate. When systematic error is minimized, error usually results from sampling variability.

Since error in small-sample DIF estimation can be viewed as deviations from the population DIF estimate, one possible way to reduce this error is to use a model to estimate the population distribution of an observed score distribution. Then the population distribution can be used in place of the observed score distribution to perform DIF analysis. Since DIF methods match examinees at each raw score level, it often becomes difficult to find data to match at each raw score level in small samples. Therefore, raw scores, which have no data to match, cannot be included in the calculation of the DIF statistic. Smoothing may offer a solution by putting

nonzero frequencies at raw score points with no data and may result in an overall improvement in the estimation of the DIF statistic.

One way to estimate the population distribution is to use smoothing techniques, which are often used to remove irregularities in observed score distributions. Log-linear models (Holland & Thayer, 1987) provide useful smoothing techniques that allow the user to specify the number of moments of the observed distributions to be preserved in the smoothed distributions. This ensures that certain basic properties of the observed data are retained in the smoothed data. If the observed sample is small, then the model may preserve only the mean and standard deviation of the observed distribution. If the observed sample is large, more moments such as the skewness and kurtosis of the observed distribution can be preserved. The log-linear method fits a model of the following form to a score distribution:

$$\log_e(m_k) = \alpha + \beta_1 k^1 + \beta_2 k^2 + \beta_3 k^3 ..... + \beta_d k^d \qquad (1)$$

where the test-score range on test *X* ranges from $k = 0$ to the maximum possible test score, k indexes the test scores, $k = 1$ to the maximum possible score + 1, n_k is the observed frequency at test score k, m_k is the smoothed frequency at test score k, d specifies the number of moments to be preserved, and α is a constant that forces $\sum_{k=0}^{K} m_k = \sum_{k=0}^{K} n_k = N$. Readers interested in the details of the log-linear smoothing method are directed to Holland and Thayer (1987), who presented a thorough description of this model including algorithms for estimation, properties of the estimates, and applications to fitting test score distributions.

Livingston (1993) used the above idea of using smoothed versus observed score distributions in estimating equating relationships in small samples and found positive results. In a DIF context, Douglas, Stout, and DiBello (1996) used a kernel-smoothed[1] version of the simultaneous item bias test (SIBTEST) procedure of Shealy and Stout (1993). They found that the smoothed version as compared to the regular version resulted in increased efficiency of local DIF (i.e., DIF observed for a specific range of the ability or score scale) estimation and improved the statistical power for detecting DIF items when the DIF hypothesis was highly local. Moreover, the smoothed version reduced noise and provided better local DIF interpretation. Although the Douglas et al. smoothed procedure was not used in the context of small-sample

DIF, the idea of smoothing may be beneficial in small-sample DIF estimation, as smoothing may help reduce noise in small samples and produce more accurate DIF estimates.

The purpose of the current study is to examine whether log-linear smoothing of score distributions improves the accuracy of DIF estimation in small samples. Although a number of different statistical methods for estimating DIF have been recommended (Angoff & Ford, 1973; Dorans & Kulick, 1986; Holland & Thayer, 1988; Shealy & Stout, 1993), this study focuses on SIBTEST.

Overview of the SIBTEST DIF Detection Procedure

SIBTEST is a nonparametric method, which was developed as an extension of Shealy and Stout's (1993) multidimensional model for DIF. In the SIBTEST framework, DIF is conceptualized as a difference between the probabilities of selecting a correct response for examinees with the same levels of the latent attribute of interest (θ). This difference, when found, is attributable to different amounts of nuisance abilities (i.e., abilities not related to the construct being measured) that influence the item-response patterns. The statistical hypothesis tested by SIBTEST is:

$$H_0 : B(T) = P_R(T) - P_F(T) = 0$$
$$Vs.$$
$$H_1 : B(T) = P_R(T) - P_F(T) \neq 0,$$

where $B(T)$ is the difference in probability of a correct response on the studied item for examinees in the reference (or advantaged) and focal (or disadvantaged) groups matched on true score; $P_R(T)$ is the probability of a correct response on the studied item for examinees in the reference group with true score T ; and $P_F(T)$ is the probability of a correct response on the studied item for examinees in the focal group with true score T . With the SIBTEST procedure, matching on target ability is done by matching on total test score or another score believed to validly measure the target ability (i.e., the dimension best measured by the reported test score), such as total score with certain items removed that are believed to be DIF-producing.

An important feature of SIBTEST is its use of a regression-correction method to match examinees from the two groups at the same true score levels (instead of observed score levels) so as to compare their performances on the studied items. The regression correction used by SIBTEST adjusts the mean (over examinees) suspect item scores to their unbiased estimates. It

does this by first computing a regression equation for each group (see Crocker & Algina, 1986, p. 147) that estimates true score on the valid subtest as a function of valid subtest observed score. This function is computed separately for each group using the estimated means, variances, and reliabilities of scores on the valid subtest. Then $B(T)$ is estimated using $\hat{B}_{UNI}$, which is the weighted sum of the differences between the proportion-correct true scores on the studied item for examinees in the two groups across all score levels. The weighted mean difference between the reference and focal groups on the studied subtest item or bundle across the k score levels in given by:

$$\hat{\beta}_{\text{UNI}} = \sum_{k=0}^{K} p_k d_k \,, \tag{2}$$

where p_k is the proportion of focal-group examinees at score level k, and d_k is the difference in the adjusted means (using regression correction) on the studied subtest item or bundles of items for the reference and focal groups, respectively, at score level k.

SIBTEST provides an overall statistical test and a measure of the effect size ($\hat{B}_{UNI}$) for each item. However, the current study will focus only on the recovery of the $\hat{B}_{UNI}$ statistic in small samples. Recovering statistical significance of the DIF estimate seems less useful in the current context because the purpose of this study is to estimate DIF in small samples, which may reduce the power of the statistical test to flag DIF items. Nevertheless, recovering statistical significance is of methodological interest and can be examined in future studies.

For classifying the SIBTEST effect sizes, guidelines first recommended by Dorans (1989) for interpreting standardized *p*-values were used. Since both the standardization and SIBTEST methods are essentially based on a measure of the same construct (i.e., the total test score), using Dorans' criteria in the SIBTEST context seemed reasonable. Nandakumar (1993) and L. Roussos (personal communication, September 15, 2005) have also recommended using Dorans' criteria for interpreting SIBTEST effect sizes and, according to Gierl and Bolt (2001), these guidelines have conceptual appeal (i.e., negligible difference is, on average, less than 1/20 of a score point between two groups, and a large difference is, on average, 1/10 of a score point or more between two groups) and some empirical support (Dorans, 1989, p. 226). Therefore, using Dorans' criteria, absolute values of the beta statistic between .000 and .050 indicate small

DIF, between .050 and .100 indicate moderate DIF, and .100 and above indicate large DIF. Items exhibiting moderate DIF should be inspected to ensure that no possible effect is overlooked. Items with large DIF are unusual and are often the ones for which test developers find content explanations for the DIF. Therefore, they should be examined very carefully (Dorans, 1989; Dorans & Holland, 1993). The term *DIF estimate* will be used instead of *beta estimate* in the remainder of the paper.

Therefore, the purpose of the current study is to examine whether log-linear smoothing of observed score distributions in small samples results in more accurate DIF estimates using the SIBTEST DIF detection method. The SIBTEST procedure described earlier is unchanged. The only difference between the smoothed and raw versions of SIBTEST is that the smoothed version will use smoothed instead of raw data. DIF estimates from several small- and moderate-sample-size conditions using raw versus smoothed data will be compared to a criterion (explained in a latter section) to determine the effect of smoothing on DIF estimation.

Method

Test Data and Analytical Procedure

The study used test data from a large-scale teacher certification test administered in 31 states. This test consists of 45 items, measures basic proficiency in writing, and is used for entrance into teaching programs. Although the test is administered in both paper-and-pencil and computerized formats, only the data from the paper-and-pencil tests were used in this study. The total data consisted of assessment results for 7,216 examinees. Of these, 6,208 examinees were in the reference group (i.e., White test takers) and 1,053 examinees were in the focal group (i.e., African American test takers). The White versus African American comparison was used because the sample size in both these groups were reasonably large enough to derive accurate criterion DIF estimates (described below). For other DIF comparisons, such as between male and female test takers or between White and Asian test takers, the sample sizes in either the reference or the focal groups were not sufficiently large enough for deriving an accurate criterion estimate. Furthermore, comparison between male and female test takers showed no gender DIF and was therefore not included in the study.

Deriving a Criterion

The current study was designed such that the population DIF estimates (or at least a close proxy of it) were known. To get the population DIF estimates, DIF analysis was conducted on the total available data, which included 44 items and 7,216 examinees. For each suspect item, the remaining items were used as the matching subtest. The current analysis includes 44 instead of 45 items because one item did not meet ETS fairness standards and therefore was not scored. The DIF estimates from the total sample were used as the criterion to which the small-sample DIF estimates were compared. Using the interpretive guidelines described earlier, items were flagged as small, moderate, or large DIF in the teacher certification test. Once the criterion DIF estimates were calculated, the study was conducted in four steps.

Step 1. Seven sample-size conditions for the focal versus reference groups were investigated in the ratio of 50/300, 75/300, 100/300, 150/300, 200/300, 300/700, and 700/2,100. Since the number of examinees in the focal group has been found to be considerably smaller than the number of examinees in the reference group for typical samples of this test, it was decided to investigate sample-size conditions where the ratio of the focal group and reference group examinees somewhat reflected this trend. It should also be noted that although the last two conditions are not necessarily considered as small-sample-size conditions, they were included in the study with the intent to show that the benefits of smoothing are the maximum for small samples and negligible for moderate and large samples (see Livingston, 1993). Within each sample-size condition, 100 samples were randomly drawn from the total data with replacement.

Step 2. DIF analysis using SIBTEST was conducted under the seven sample-size conditions. For each studied item, DIF analysis was conducted for each of the 100 samples in the first condition (i.e., the 50/300 sample-size condition). Then the 100 small samples were smoothed using the log-linear smoothing procedure. Specifically, the frequency of the rights and wrongs of the studied item at each raw score level of the matching subtest in both the reference and focal groups were smoothed. After some tryout analyses it was decided to preserve three moments (i.e., $d = 3$) in the smoothing, since it provided a good fit for most sample-size conditions (i.e., the likelihood ratio chi-square statistics were close in value to the models' degrees of freedom for most considered items and sample sizes). Thus, the log-linear smoothing equation simplified to

$$\log_e(m_k) = \alpha + \beta_1 k^1 + \beta_2 k^2 + \beta_3 k^3, \quad (3)$$

where the test score range on test *X* goes from $k = 0$ to the maximum possible test score (*K*), k represents the test scores, $k = 1$ to the maximum possible score + 1, n_k is the observed frequency at test score *k*, m_k is the smoothed frequency at test score *k*, and α is a constant that forces $\sum_{k=0}^{K} m_k = \sum_{k=0}^{K} n_k = N$. β_1, β_2 and β_3 are estimated when the model is fit. They are estimated so that the mean, variance, and skew in the observed distribution are preserved in the smoothed distribution:

$$\sum_{k=0}^{K} k^1 (\frac{m_k}{N}) = \sum_{k=0}^{K} k^1 (\frac{n_k}{N}) \text{ for the mean,}$$

$$\sum_{k=0}^{K} k^2 (\frac{m_k}{N}) = \sum_{k=0}^{K} k^2 (\frac{n_k}{N}) \text{ for the variance, and}$$

$$\sum_{k=0}^{K} k^3 (\frac{m_k}{N}) = \sum_{k=0}^{K} k^3 (\frac{n_k}{N}) \text{ for the skewness.}$$

After the raw data was smoothed, DIF analysis was performed for the 100 smoothed samples. This step was repeated for all studied items and remaining sample-size conditions. Thus, for each studied item, there were 700 raw and 700 smoothed DIF runs (i.e., 100 replications *X* 7 sample-size conditions).

Step 3. Both the raw DIF estimates and the smoothed DIF estimates were compared to the criterion estimates derived from the total sample to evaluate whether smoothing resulted in more accurate DIF estimates. While comparing the raw and smoothed DIF estimates with the criterion DIF estimate, both variability and accuracy of the DIF estimates were evaluated using two statistics. The root-mean-squared deviation (RMSD) of the raw DIF estimates and the smoothed DIF estimates from the criterion DIF estimate was used to measure the overall variability of a DIF estimation procedures (i.e., unsmoothed versus smoothed). To compute this statistic, let y represent the criterion DIF estimate for a particular item, let $\hat{y}_r$ represent the DIF estimate for the

item using raw data, and let $\hat{y}_s$ represent the DIF estimate for the item using smoothed data. Thus. for the raw data, the RMSD across the 100 replications within each condition is defined as:

$$RMSD = \sqrt{\frac{1}{100}\sum_{j=1}^{100}(\hat{y}_r - y)^2} \tag{4}$$

The same equation is used to calculate the RMSD for the smoothed data by replacing the $\hat{y}_r$ term with the $\hat{y}_s$ term. The smaller the RMSD for smoothed versus raw DIF estimates within each sample-size condition, the less variable is the estimate.

To measure the accuracy of a DIF estimation procedure, an overall bias statistic was also calculated using the formula stated below. The same equation is used to calculate the bias for the smoothed data by replacing the $\hat{y}_r$ term with the $\hat{y}_s$ term:

$$Bias = \left(\frac{1}{100}\sum_{j=1}^{100}\hat{y}_r\right) - y \tag{5}$$

Although the bias statistic cancels out positive and negative differences around the criterion, it is helpful in summarizing whether, on an average, the smoothed or raw estimates within each sample-size condition deviates from the criterion in either the positive or negative direction. An overall bias statistic that is close to zero would indicate a more accurate DIF estimation.

Step 4. Finally, the percentage of time that the smoothed and the raw estimates fell within 0.1 units (i.e., beta units, which is the DIF unit of SIBTEST) above or below the criterion estimate was calculated. Since the guidelines for DIF interpretation as described earlier suggests that a DIF estimate greater than 0.1 for an item indicates large DIF and should be examined very carefully, it seemed reasonable to create the 0.1 band above and below the criterion DIF estimates. Estimates that fell outside this band would be considered too large to ignore.

Results

The criterion DIF analysis conducted on the large sample (i.e., the total available data for the particular test form) flagged 6 items with potential for DIF (i.e., Items 1, 4, 9, 22, 28, and 35).

Following the DIF-interpretation guidelines describe earlier, two items showed large DIF and four items showed moderate DIF in the White/African American comparison. These six items and two other items showing small DIF were used as the studied items, and the small-sample DIF estimates for these items using the raw and smoothed data were compared to their criterion DIF estimates to evaluate the benefits of smoothing the data. The value of the DIF estimates, proportion correct, and point biserial (or discrimination parameter) for all eight items are shown in Table 1.

Table 1

Criterion DIF Estimates, Proportion Correct, and Discrimination

Items	Sample size (focal vs. reference)	Proportion correct	Point-biserial	DIF (beta) estimate
1	1,053/6,208	0.761	0.309	–0.110 L
4	1,053/6,208	0.882	0.349	0.086 M
7	1,053/6,208	0.779	0.193	–0.012 S
9	1,053/6,208	0.775	0.294	0.113 L
22	1,053/6,208	0.442	0.397	0.059 M
28	1,053/6,208	0.557	0.215	–0.069 M
35	1,053/6,208	0.239	0.266	0.056 M
42	1,053/6,208	0.588	0.429	–0.022 S

Note. S, M, and L denote small, moderate, and large DIF levels where S indicates beta values between 0.000 and 0.050, M indicates beta values between 0.050 and 0.100, and L indicates beta values greater than 0.100.The RMSD and bias statistics were calculated across replications for all combinations of items and sample sizes. These values are presented in Tables 2 and 3 and are also summarized in graphs presented in Figures 1 and 2. Finally, for each item, the actual value of the 100 raw and 100 smoothed estimates in each sample-size condition are summarized in box plots shown in Figure 3.

RMSD and Bias Results

As seen in Table 2 and Figures 1 and 2, the RMSD results for the eight items suggests that for these items the RMSD was generally lower for both the raw and the smoothed estimates for larger (as compared to the smaller) sample-size conditions. The RMSD for the smoothed and

raw DIF estimates showed slightly inconsistent results for different items. For Items 1, 4, 7, 9, 28, 35, and 42, the RMSDs for the smoothed estimates were considerably smaller than the RMSDs for the raw estimates in the smaller sample-size conditions. However, as the sample size got larger, the difference between the RMSDs for the raw and smoothed estimates became smaller. This is not surprising, since smoothing procedures are expected to bring the greatest benefits for the small-sample-size conditions (Livingston, 1993). The RMSD for Item 22 showed a different pattern. It was lower for the smoothed estimates (as compared to the raw estimates) for the first three sample-size conditions. However, the reverse was true for the last four sample-size conditions (although it should be noted that the difference in the RMSDs for the raw and smoothed estimates in these four sample-size conditions was very small). Overall, these results seem to suggest that smoothing was beneficial for most items especially in the small-sample-size conditions.

As seen in Table 3 and Figures 3 and 4, the bias values for the smoothed versus the raw data for the eight items showed mixed results. For Items 7 and 28, the bias statistic was lower (i.e., closer to zero) for the smoothed as compared to raw estimates in all sample size conditions. For Items 1, 4, and 35, the bias values were lower for the smoothed as compared to the raw estimates for most sample-size conditions. These items showed a slightly larger bias for the smoothed as compared to the raw estimates for the largest sample-size condition (i.e., 700/2,100). However, the difference in the bias values for the raw and smoothed estimates was very small. For Items 9 and 22, the bias values were lower for the smoothed as compared to the raw estimates for the small-sample-size conditions. However, as the sample got larger, the bias for the smoothed estimates became larger than the raw estimates. Finally, the bias statistics for Item 42 showed that the bias was close to zero for both the raw and the smoothed estimates in the lower sample-size conditions (i.e., the first two conditions). The bias for the smoothed estimates was always lower than the bias of the raw estimates for the remaining five sample-size conditions. From these results, it also seems that the benefit of smoothing on DIF estimation was apparent for most items, especially in the small-sample-size conditions.

Table 2

RMSD Across the Seven Sample-Size Conditions for the Eight Items

Items	Sample-size conditions						
	50/300	75/300	100/300	150/300	200/300	300/700	700/2100
1							
R	0.0160	0.0117	0.0103	0.0063	0.0050	0.0035	0.0016
S	0.0072	0.0056	0.0049	0.0045	0.0038	0.0026	0.0013
4							
R	0.0145	0.0109	0.0072	0.0046	0.0044	0.0030	0.0015
S	0.0066	0.0057	0.0044	0.0040	0.0032	0.0029	0.0021
7							
R	0.0192	0.0148	0.0105	0.0072	0.0053	0.0039	0.0014
S	0.0081	0.0061	0.0062	0.0047	0.0040	0.0032	0.0014
9							
R	0.0149	0.0104	0.0088	0.0068	0.0048	0.0044	0.0018
S	0.0077	0.0076	0.0056	0.0052	0.0039	0.0028	0.0023
22							
R	0.0113	0.0088	0.0082	0.0060	0.0054	0.0034	0.0020
S	0.0098	0.0069	0.0077	0.0068	0.0055	0.0036	0.0033
28							
R	0.0115	0.0096	0.0080	0.0074	0.0062	0.0042	0.0021
S	0.0095	0.0076	0.0072	0.0062	0.0051	0.0038	0.0018
35							
R	0.0159	0.0117	0.0092	0.0060	0.0047	0.0034	0.0015
S	0.0075	0.0058	0.0058	0.0043	0.0041	0.0029	0.0017
42							
R	0.0110	0.0082	0.0074	0.0065	0.0066	0.0050	0.0020
S	0.0089	0.0079	0.0068	0.0055	0.0050	0.0034	0.0015

Note. R = raw estimates; S = smoothed estimates.

Table 3

Bias Across the Seven Sample-Size Conditions for the Eight Items

Items	Sample-size conditions						
	50/300	75/300	100/300	150/300	200/300	300/700	700/2100
1							
R	0.1059	0.0820	0.0569	0.0166	0.0012	0.0016	0.0005
S	–0.0158	–0.0141	–0.0112	–0.0090	–0.0061	–0.0017	0.0013
4							
R	0.0868	0.0478	0.0237	–0.0033	–0.0126	–0.0042	–0.0049
S	–0.0192	–0.0159	–0.0080	–0.0132	–0.0076	–0.0160	–0.0177
7							
R	0.1528	0.1084	0.0728	0.0372	0.0241	0.0164	0.0042
S	0.0089	0.0010	0.0023	0.0005	0.0019	–0.0046	–0.0007
9							
R	0.1072	0.0519	0.0565	0.0278	0.0168	0.0238	0.0051
S	–0.0017	–0.0193	–0.0097	–0.0070	–0.0082	–0.0059	–0.0176
22							
R	–0.0323	–0.0441	–0.0275	0.0064	0.0059	0.0018	–0.0001
S	0.0068	0.0035	0.0166	0.0186	0.0205	0.0169	0.0264
28							
R	0.0670	0.0498	0.0398	0.0301	0.0264	0.0228	0.0081
S	0.0003	0.0048	–0.0024	–0.0022	0.0010	0.0068	0.0076
35							
R	–0.0918	–0.0785	–0.0568	–0.0323	–0.0080	–0.0133	–0.0020
S	0.0119	0.0097	0.0131	0.0053	0.0106	0.0046	0.0037
42							
R	0.0000	–0.0014	0.0137	0.0307	0.0328	0.0265	0.0134
S	–0.0021	–0.0057	0.0038	0.0084	0.0079	0.0006	–0.0012

Note. R = raw estimates; S = smoothed estimates.

Table 4

Percentage of Raw and Smoothed Estimates Falling Within 0.1 (+/-) Units of Criterion Estimate

Items	Sample-size conditions						
	50/300	75/300	100/300	150/300	200/300	300/700	700/2100
1							
R	71	94	93	**98**	99	100	100
S	**88**	94	**98**	97	**100**	100	100
4							
R	53	72	82	98	96	100	100
S	**86**	**91**	**99**	**99**	**99**	100	100
7							
R	31	47	65	83	95	99	100
S	**84**	**91**	**88**	**95**	**99**	99	100
9							
R	39	68	74	88	98	100	100
S	**79**	**80**	**94**	**95**	98	100	100
22							
R	60	74	78	**90**	95	100	100
S	**69**	**85**	**82**	87	95	100	100
28							
R	62	71	80	84	90	98	100
S	**77**	**80**	**88**	**88**	**94**	**99**	100
35							
R	51	60	70	88	95	100	100
S	**84**	**91**	**90**	**98**	**99**	100	100
42							
R	67	27	85	88	86	96	100
S	**72**	**66**	**89**	**90**	**95**	**100**	100

Note. R = raw estimates; S = smoothed estimates. For each raw versus smoothed comparison, bold digits indicate a larger number.

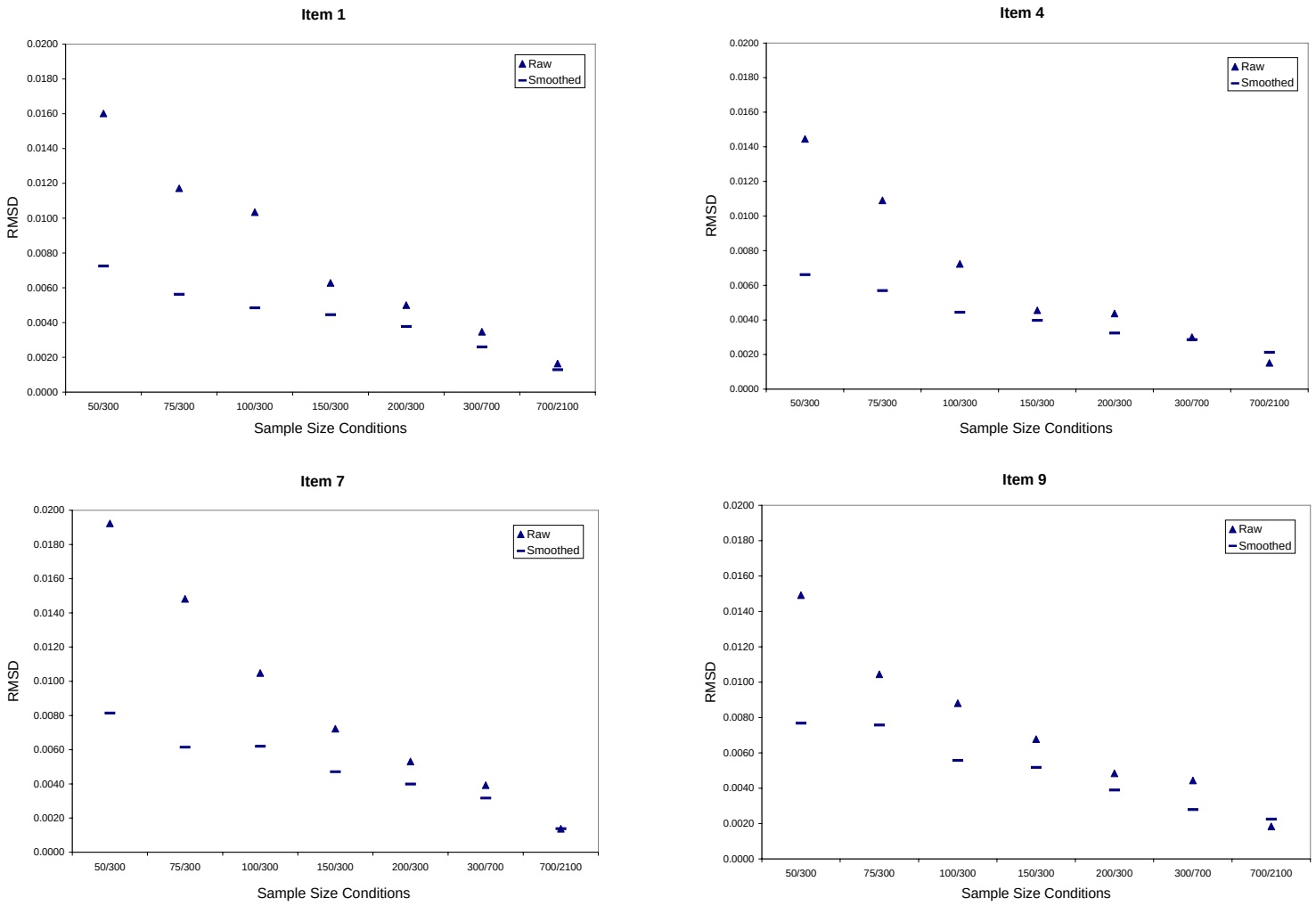

Figure 1. **RMSD for DIF estimates for smoothed and raw data for Items 1, 4, 7, and 9.**

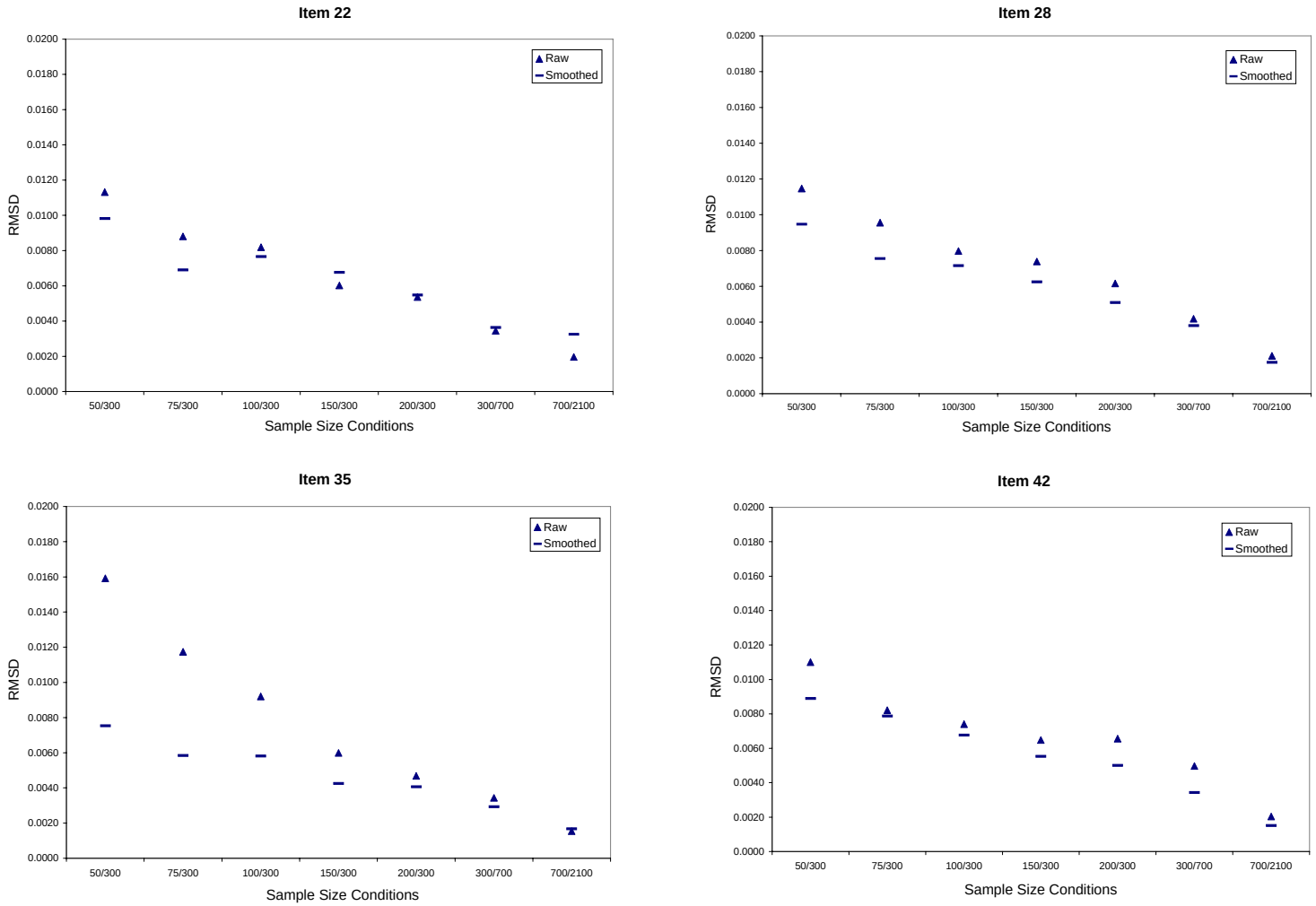

Figure 2. **RMSD for DIF estimates for smoothed and raw data for Items 22, 28, 35, and 42.**

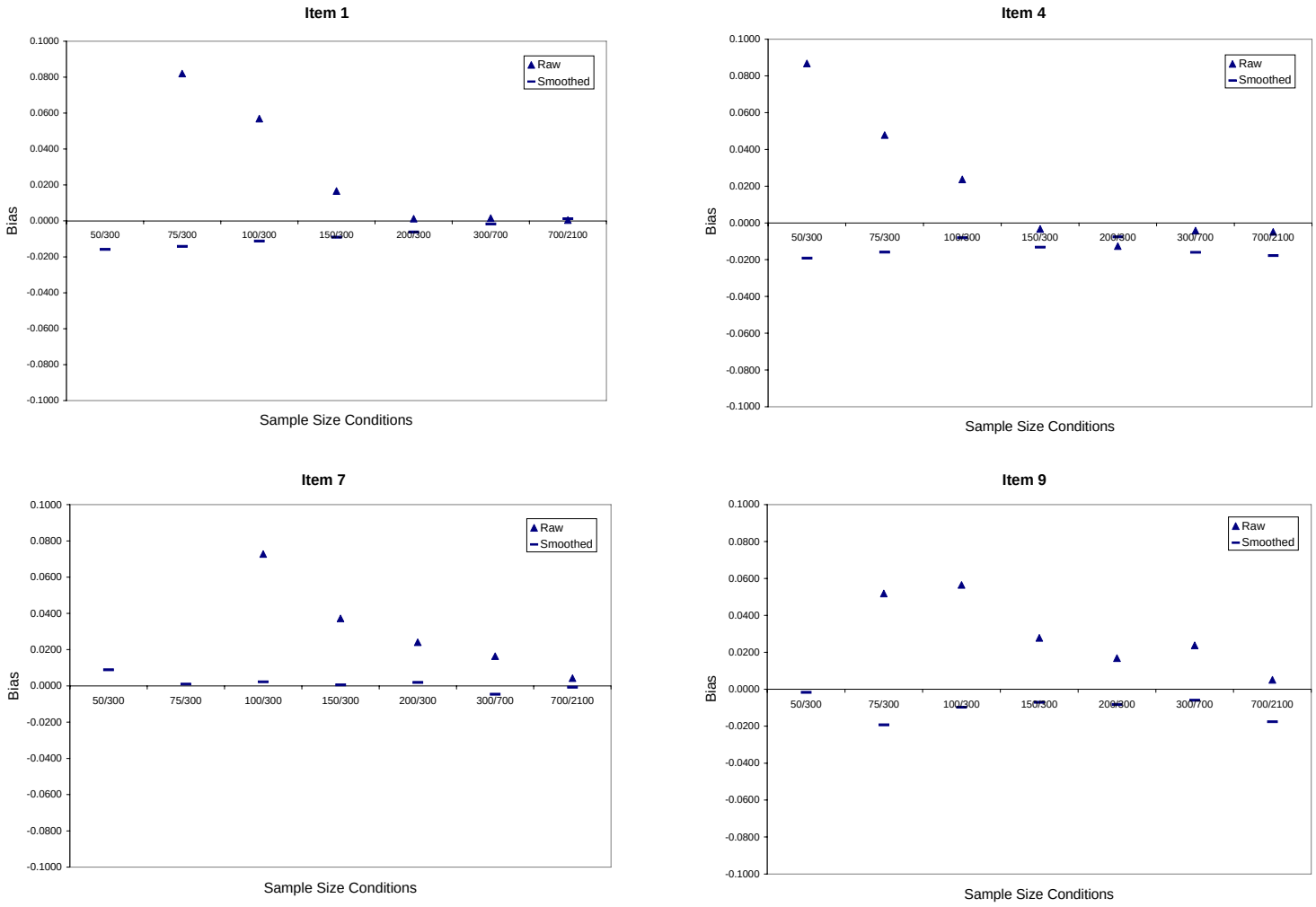

Figure 3. **Bias for DIF estimates for smoothed and raw data for Items 1, 4, 7, and 9.**

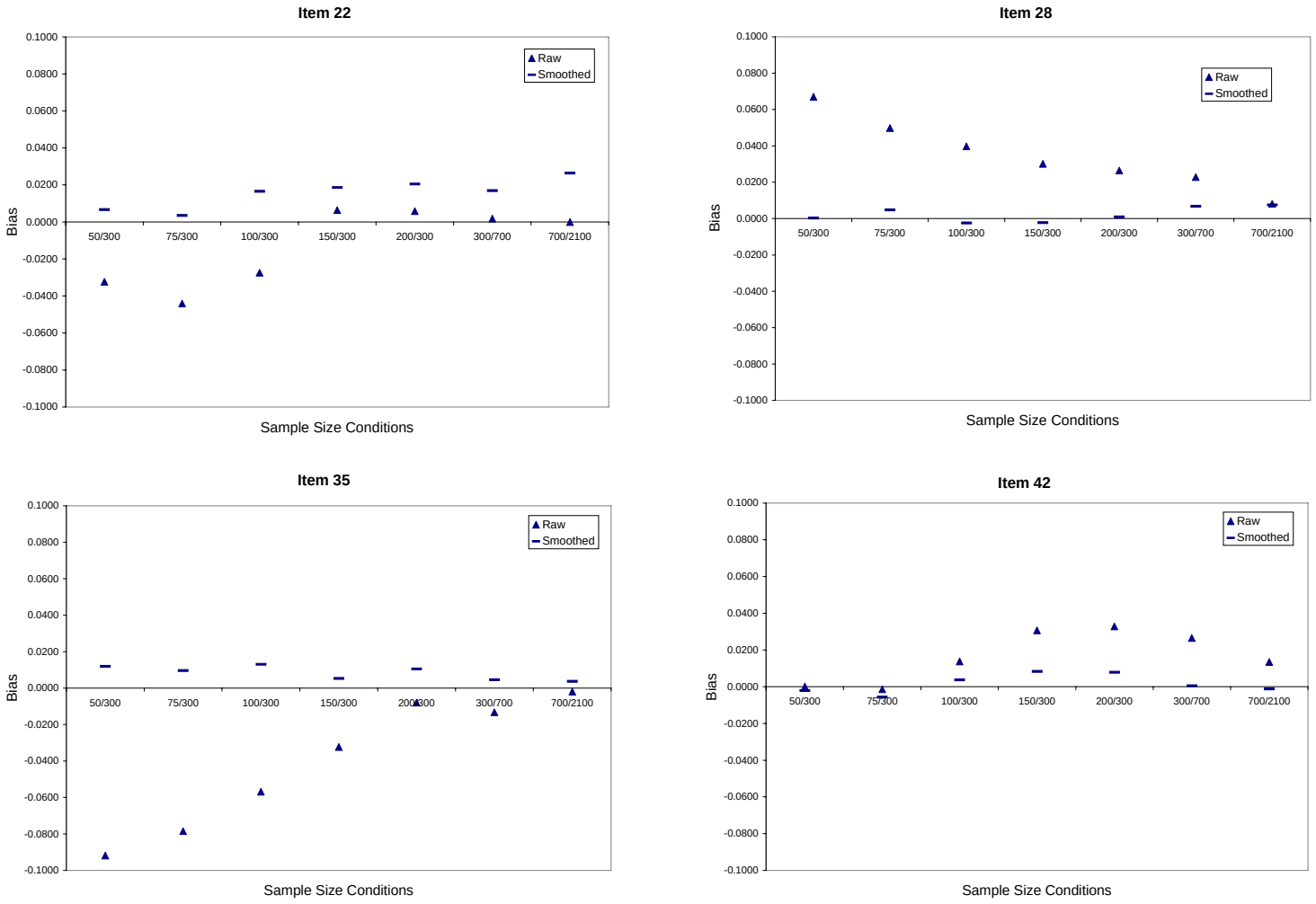

Figure 4. **Bias for DIF estimates for smoothed and raw data for Items 22, 28, 35, and 42.**

Box Plot Summary of Variability and Bias

The box plots that summarize both the variability and bias are presented in Figures 5 and 6. The variability of the estimates for all items tends to get smaller as the sample sizes become larger. Although the variability of the smoothed estimates was smaller for most items and sample-size conditions, some items show that the variability of the smoothed estimates is slightly larger than the raw estimates under some sample-size conditions (e.g., Items 22 and 28 at sample size 50/300). The median point shows how close the smoothed and raw estimates are to the criterion estimate for different items and sample-size conditions. For Items 1, 7, 9, 28, 35, and 42, the median value was closer to the criterion for the smoothed estimates as compared to the unsmoothed estimates for most sample-size conditions. For Items 4 and 22, the median value for the smoothed estimates as compared to the raw estimates was closer to the criterion estimate in the small-sample-size conditions. The reverse was true in some of the larger sample-size conditions. Like the RMSD and bias results presented earlier, these results also suggest that smoothing is beneficial in reducing variability and bias for most items, especially in the small-sample-size conditions.

Smoothed and Raw Estimates Within 0.1 Units Above or Below the Criterion

The percentage of raw and smoothed estimates falling within 0.1 units above or below the criterion are summarized in Table 4. As seen in the table, the smoothed estimates fell within this range more often than the raw estimates. This was true for all the eight items. Moreover, for these items, the difference in the percentage of time the smooth estimates fell within the range as compared to the raw estimates became smaller as the sample-size conditions became larger (i.e., there was no or negligible difference especially for the last three sample-size conditions). These results suggest that smoothing was beneficial for all items, especially in the smaller sample-size conditions.

Discussion and Conclusion

The purpose of the present study was to evaluate whether log-linear smoothing of score distributions leads to more accurate DIF estimation, especially in the case of small samples. The accuracy of the smoothed versus the raw DIF estimates within the SIBTEST framework was evaluated by comparing these estimates with criterion estimates derived from the total sample. The deviations from the criterion were assessed using the RMSD and bias statistics.

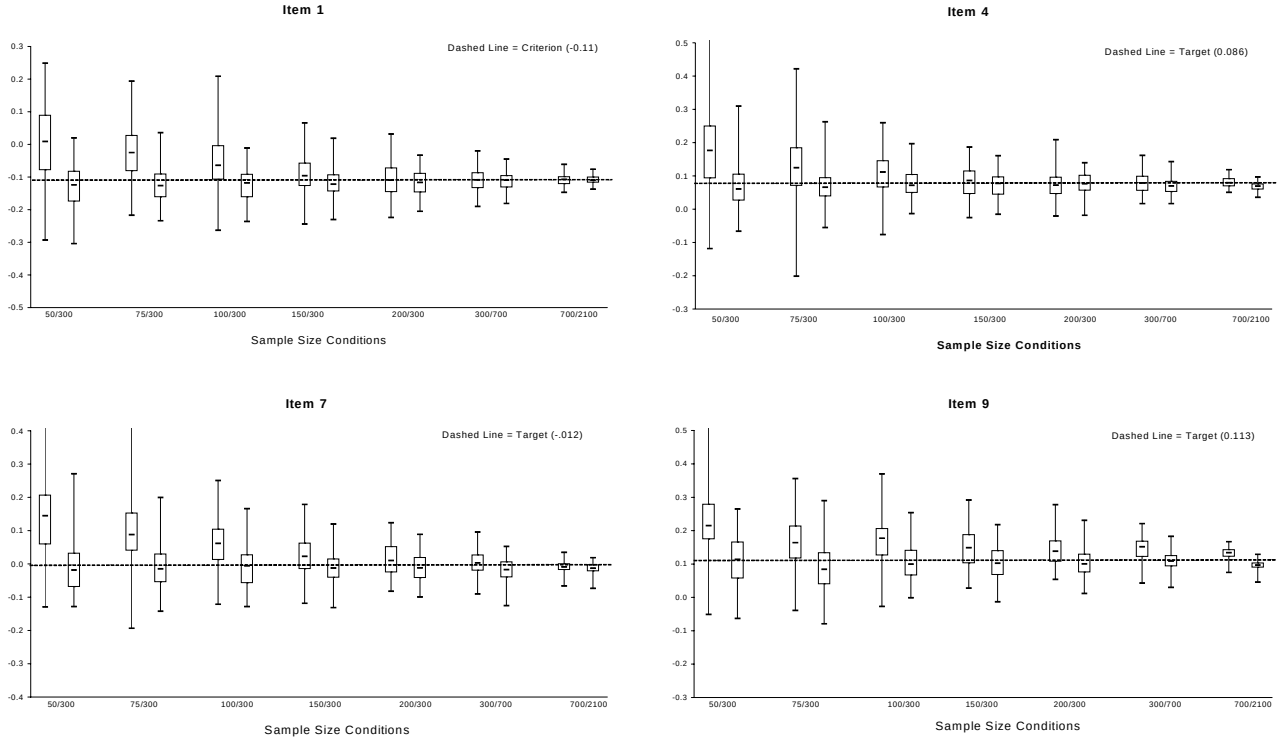

Figure 5. **Box plot summaries for smoothed and raw data for Items 1, 4, 7, and 9.**

Note. For box plots: (a) The scale width is fixed at approximately 0.4 DIF units above and below the criterion and therefore, in some cases where the small-sample DIF estimates fell outside this band, the box plots appears to be cut off; (b) under all sample-size conditions, the first and second box plots represent the raw and smoothed estimates, respectively.

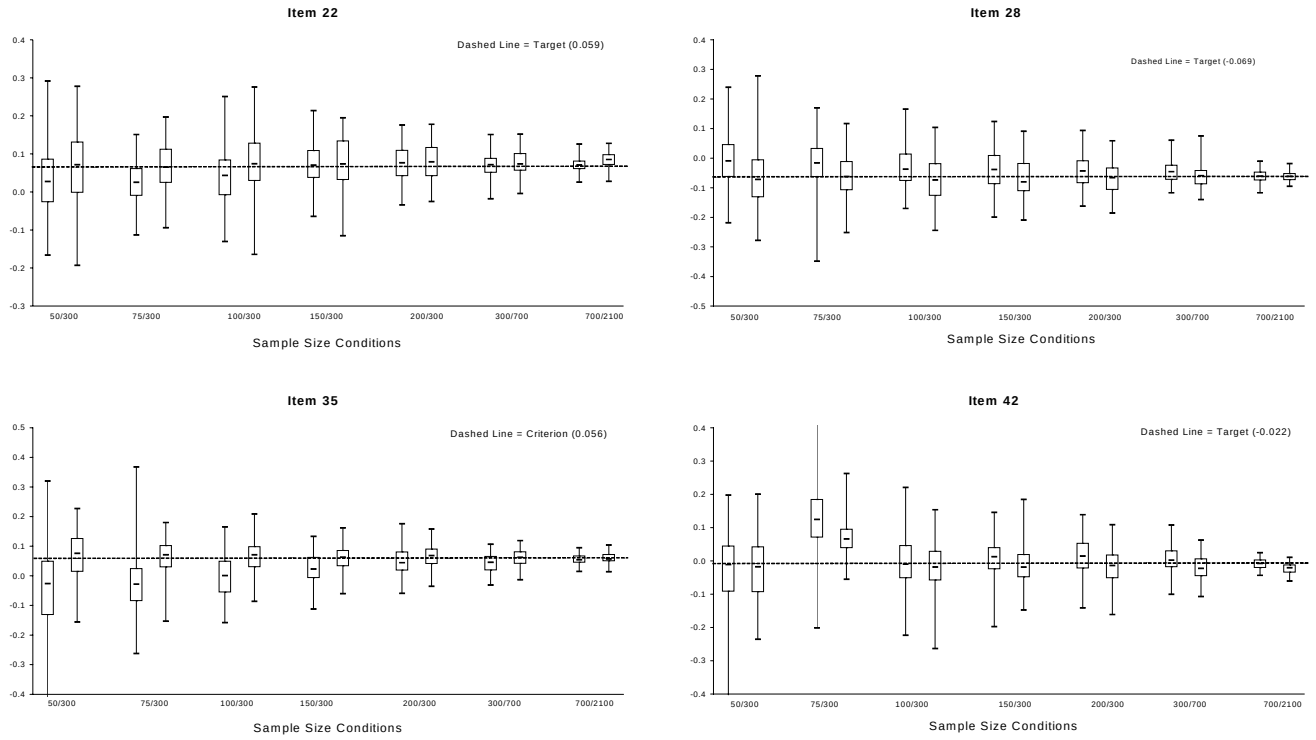

Figure 6. **Box plot summaries for smoothed and raw data for Items 22, 28, 35, and 42.**

Note. For box plots: (a) The scale width is fixed at approximately 0.4 DIF units above and below the criterion and therefore, in some cases where the small-sample DIF estimates fell outside this band, the box plots appears to be cut off; (b) under all sample-size conditions, the first and second box plots represent the raw and smoothed estimates, respectively.

In general, the RMSD results suggested, that for most items, smoothing resulted in less variability of the DIF estimates, especially in the smaller sample-size conditions. However, the benefit of smoothing for some items was unclear (i.e., RMSD for the smoothed estimates were lower than the raw estimates for some sample-size conditions but not for others). Similarly, the bias results suggested that, for most items, the smoothed estimates were quite close to the criterion, especially in the small-sample-size conditions. For some items, however, the bias in the smoothed estimates was slightly larger than the raw estimates in the large-sample-size conditions.

The box plots summarizing the results from the smoothed and raw estimates also yielded similar results, suggesting that the benefit of smoothing was apparent for most items, especially in the small-sample-size conditions. However, the benefit of smoothing for other items was unclear (e.g., Item 22).[2] Since smoothing produced the expected benefits for some items and not for others, we looked to see whether item characteristics had an effect on better estimates in the smoothed condition. For example, were the items for which the smoothed SIBTEST worked well extremely easy or difficult items? Also, any consistent pattern in the discrimination parameter for these items was examined. However, no consistent difficulty or discrimination pattern was found for these items.

Finally, the percentage of time the smoothed and raw estimates fell within 0.1 units above or below the criterion was investigated. These results showed that, in general, the smoothed estimates fell within this range more often as compared to the raw estimates for most items, suggesting that smoothing was useful in reducing the variability and increasing the accuracy of DIF estimates for all items, especially in the small-sample size-conditions.

Although the benefits of smoothing was less clear for some items, these results are somewhat encouraging overall, as they show that smoothing was useful in reducing variability and improving the accuracy of DIF estimates, especially in the small-sample-size conditions. If testing programs are conducting DIF on small samples (based on the raw data), then the smoothing procedure described in this study may provide an improvement over the current procedure.

However, many testing programs do not conduct DIF analysis for tests with small samples because of minimum sample-size requirements provided in the DIF literature. For example, Mazor, Clauser, and Hambleton (1992) suggested 200 examinees as a minimum

sample size for conducting DIF analysis. In practice, testing administrations often have subgroups with sample sizes much below this minimum requirement. Although test developers and item reviewers often rely on DIF statistics to make more informed decisions about the quality of an item, testing programs sometimes choose not to conduct DIF analysis on small samples in order to prevent test developers and item reviewers from getting DIF statistics based on small samples, which are often not stable and may, in turn, lead the test developers and/or reviewers to make an inaccurate conclusion. As evident, results of the current study does not provide a strong enough reason to suggest that these testing programs start conducting DIF on small samples under the log-linear smoothing framework. Although the smoothed estimates show some improvement over the raw estimates, they are still not close to the criterion in several cases.

Future Research

A strong need remains to conduct DIF on small samples. As Parshall and Miller (1995) pointed out, state boards and licensing agencies often make contractual requirements for DIF analyses to be performed, regardless of the statistical appropriateness of sample size. Similarly, to make more informed decisions, test developers and item writers use DIF statistics when conducting substantive reviews of test items. DIF statistics most likely help in confirming or rejecting their substantive judgments. Therefore, future research should continue investigating methods for conducting DIF on small samples.

Future research may focus on different approaches than those followed in the current study. For example, smoothing techniques other than log-linear smoothing can be used. During the initial phase of this study, kernel smoothing (Douglas, Stout, & DiBello, 1996) was used as a second method of smoothing in addition to log-linear smoothing. However, after some preliminary analyses with kernel smoothing, it was decided to drop this smoothing method for two main reasons. First, results from kernel smoothing were very similar to results from log-linear smoothing for the first 30 replications in the seven sample-size conditions. Second, using kernel smoothing in the SIBTEST framework took a much longer time to conduct the DIF analyses (e.g., it took more than 24 hours to run the smoothed DIF program for the seven items on a standard personal computer). For testing programs with strict timelines, usually a day or two is allotted for DIF analysis before equating and score reporting can be completed. In case of kernel smoothing, it would take much longer than two days to get DIF estimates for several new

forms introduced in a testing administration. Therefore, this smoothing procedure was dropped from the current study for practical reasons. However, kernel smoothing is still of methodological interest and should be researched further to examine its effect on small-sample DIF estimation and to assess whether it produces similar results as compared to log-linear smoothing over larger replications.

The current study examined the effect of smoothing in the SIBTEST framework. However, other methods of DIF detection (i.e., standardization and Mantel-Haenzel) also exist and are frequently used by testing programs to conduct DIF analysis. Therefore, future research should also investigate the effect of smoothing techniques on other DIF-estimation procedures.

Finally, in the current study, the true values of DIF were estimated using the total available data. Although one can be fairly confident that the true values generated using the total available data would be a good proxy of the true DIF value, it is possible to get a slightly different value if a different sample were used. Therefore, it may be useful to investigate the effect of smoothing in small-sample DIF estimation under simulated conditions. A simulated study may provide a better control in the generation of criterion DIF values. Since one would generate a true value based on a theoretical framework, the true value would not differ if we generated it again using the criteria followed previously.

References

Angoff, W. H., & Ford, S. F. (1973). Item-race interaction on a test of scholastic aptitude. *Journal of Educational Measurement*, *10*, 95–106.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice, 17*(1), 31–44.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Belmont, CA: Wadsworth Group/Thomson Learning.

Dorans, N. J. (1989). Two new approaches to assessing differential item functioning: Standardization and the Mantel-Haenszel method. *Applied Measurement in Education, 2*(3), 217–233.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35–66). Hillsdale, NJ: Lawrence Erlbaum.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing the unexpected differential item functioning on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23,* 355–368.

Douglas, J. A., Stout, W, & DiBello, L. V. (1996). A kernel-smoothed version of SIBTEST with applications to local DIF inference and function estimation, *Journal of Educational and Behavioral Statistics*, *21*, 333–363.

Gierl, M. J., & Bolt, D. M. (2001). Illustrating the use of nonparametric regression to assess differential item and bundle functioning among multiple groups. *International Journal of Testing, 1*(3 & 4), 249–270.

Holland P. W., & Thayer D. T. (1988). Differential item performance and Mantel- Haenszel. In H. Wainer & H. Braun (Eds.), *Test Validity*. Hillsdale, NJ: Lawrence Erlbaum.

Holland, P. W., & Thayer, D. T. (1987). *Notes on the use of the log-linear models for fitting discrete probability distributions*. Technical Report No. 87-79. Princeton, NJ: ETS.

Livingston, S. A. (1993). Small-sample equating with log-linear smoothing. *Journal of Educational Measurement,* 30, 23–29.

Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1992). The effect of sample size on the functioning of the Mantel-Haenszel statistic. *Educational and Psychological Measurement, 52,* 443–451.

Nandakumar, R. (1993). Simultaneous DIF amplification and cancellation: Shealy–Stouts test for DIF. *Journal of Educational Measurement, 16,* 159–176.

Parshall, C. G., & Miller, T. R. (1995). Exact versus asymptotic Mantel-Haenszel DIF statistics: A comparison of performance under small-sample conditions. *Journal of Educational Measurement, 32,* 302–316.

Shealy, R., & Stout, W. F. (1993). A model-based standardization approach that separates true bias/DIF from group differences and detects test bias/DIF as well as item bias/DIF. *Psychometrika, 58*, 159–194.

Notes

[1] Kernel smoothing is a technique that borrows information from neighboring scores to get a more stable estimate at each score level.

[2] Since the results for Item 22 were puzzling, additional analyses were conducted to examine factors that may have caused these results. These analyses and results, and their interpretation, are presented in the appendix.

Appendix

The following analyses were conducted to further explore the ambiguous results for Item 22. These analyses point to several explanations for the large-sample differences in the raw and smoothed SIBTEST results of Item 22, and illustrate how the smoothing and SIBTEST procedures worked together for a somewhat unique item. They also allow for a detailed consideration of how the regression correction worked on raw and erratic data. To review, the SIBTEST procedure is based on the following steps:

1. Obtain reference and focal conditional average item scores for the total levels of the matching variable.
2. Apply a regression correction to estimate reference and focal average item scores at common levels of the true score (which is linearly related to the observed score of the matching variable).
3. Compute $\hat{B}_{UNI}$ as a weighted average of the difference in regression-corrected average item scores across all K levels of the (transformed) matching variable.

To make use of loglinear smoothing, Step 1 was modified so that the conditional average item scores would be computed from smoothed, rather than raw frequencies of the distributions of correct and incorrect responses to the studied item.

Figures A1–A4 plot the four smoothed (circles) and raw (squares) distributions used in Step 1 to compute conditional reference and focal average item scores for the studied item (Item 22) for the total population (6,208 reference examinees and 1,053 focal examinees). These four distributions are of the reference examinees who got Item 22 correct (Figure A1), reference examinees who got Item 22 incorrect (Figure A2), focal examinees who got Item 22 correct (Figure A3), and focal examinees who got Item 22 incorrect (Figure A4).

One of the possible explanations for the differences between the large-sample smoothed and raw SIBTEST estimates was that the smoothing models did not fit the data. Figures A1–A4 do not show serious model misspecifications (i.e., the raw and smoothed frequencies are not systematically different). Figures A3–A4 indicate that the raw focal frequencies, which are based on a smaller total sample size than the raw reference frequencies, have substantial unsystematic noise. The four likelihood ratio chi-square statistics, overall measures of model misfit, are insignificant ($\chi^2 = 36.67$ for the reference correct distribution; $\chi^2 = 46.11$ for the reference

incorrect distribution; $\chi^2 = 45.79$ for the focal correct distribution; and $\chi^2 = 44.69$ for the focal incorrect distribution, all on 40 degrees of freedom, $p > .10$).

The smoothed and raw frequencies in Figures A1–A4 were used to compute smoothed and raw conditional average item scores [= frequency correct/(frequency correct + frequency incorrect)]. The reference (circles) and focal (squares) conditional average item scores are plotted in Figures A5 (raw) and A6 (smoothed). The raw average item scores that are plotted are those that are actually used in the computation of the SIBTEST $\hat{B}_{UNI}$: They are not at the minimum or maximum valid subtest scores, the reference frequency is at least 1, the focal frequency is at least 1, and the item score variances are greater than zero (Shealy & Stout, 1993). The focals' average item scores are of particular interest because they are not monotonically increasing across the scores of the matching variable, and also because of their high degree of fluctuation. The focals' raw average item scores are even greater than the references' raw average item scores at subtest scores of 26, 28, 29, and 31. Figure A4 shows that the frequencies of focal examinees who got Item 22 incorrect at subtest scores of 26, 28, 29, and 31 were abnormally low relative to the raw frequencies at adjacent scores and also relative to the smoothed frequencies at the same scores.

The smoothed average item scores extend to the entire score range of the valid subtest's score range (Figure A6), while the raw average item scores do not (Figure A5). The smoothing fills in frequencies where there were originally zero frequencies. This extrapolation is based on models that are intended to fit the observed data well, but is a series of guesses that extend beyond the actual data. One result is that the smoothed $\hat{B}_{UNI}$ was relatively large compared to the raw $\hat{B}_{UNI}$ because the extrapolated average item score differences at the extrapolated valid subtest scores were greater than those at the observed score levels.

The regression correction is applied to the conditional average item scores and is also to some extent based on the average item scores. The regression correction as proposed by Shealy and Stout (1993) is as follows:

$$\bar{Y}^{*}_{gk} = \bar{Y}_{gk} + \frac{\bar{Y}_{g,k+1} - \bar{Y}_{g,k-1}}{V_{g,k+1} - V_{g,k-1}} \left(\frac{V_{Rk} + V_{F,k}}{2} - V_{gk} \right), \tag{5}$$

where $\bar{Y}_{gk}$ is the unadjusted average item score for group *g* at valid subtest score k and V_{gk} is the estimated true score for group *g* at score *k* based on the reliability and mean of the valid subtest *X* in group *g*,

$$V_{gk} = \frac{\bar{X}_g + rel_{Xg}(k - \bar{X}_g)}{K}. \quad (6)$$

The goal of the regression correction is to remove groups' valid subtest distribution differences from the conditional average item scores. For Item 22, the means on the valid subtest were 22.76 and 18.13 in the reference and focal groups, respectively. This mean difference was in the direction that is portrayed in Shealy and Stout (1993, p. 171–172) and is what is often the case in DIF studies, the focal group is less able on the matching test than the reference group. Shealy and Stout proposed their regression correction because of a worry that mean differences on the observed scores of the matching variable would inflate the differences in conditional average item scores, and therefore suggest DIF where no DIF exists.

In (5), the regression correction to the observed average item score is based on $\frac{\bar{Y}_{g,k+1} - \bar{Y}_{g,k-1}}{V_{g,k+1} - V_{g,k-1}}$ and $\frac{V_{Rk} + V_{F,k}}{2} - V_{gk}$. The first term is an estimate of the rate of change in a groups' average item score (numerator) as a function of the increase in the latent trait score (denominator). Since it is generally believed that item-response functions are monotonically increasing (Shealy & Stout, 1993, p. 171) because more able examinees ought to score higher than less able examinees on individual items, $\frac{\bar{Y}_{g,k+1} - \bar{Y}_{g,k-1}}{V_{g,k+1} - V_{g,k-1}}$ should be positive for increases in *k*. The sign of the $\frac{V_{Rk} + V_{F,k}}{2} - V_{gk}$ term will be negative for a more able reference group $\left(\frac{V_{Rk} + V_{F,k}}{2} < V_{Rk}\right)$, so that the reference group's conditional average item score ($\bar{Y}_{Rk}$) should ultimately be reduced by the regression correction. The sign of $\frac{V_{Rk} + V_{F,k}}{2} - V_{gk}$ will be positive

for a less able focal group $\left(\frac{V_{Rk}+V_{F,k}}{2} > V_{Fk}\right)$, so that the focal groups' conditional average item score ($\bar{Y}_{Fk}$) should ultimately be increased by the regression correction.

The functioning of Item 22 for focal examinees is inconsistent with the SIBTEST regression correction in two important respects. First, the focal examinees' average item scores are not monotonically increasing. Second, focal average item scores are greater than the reference average item scores for some levels of the matching variable (i.e., if average item scores actually contain reference and focal differences on the matching variable, then a less-able focal group should not have average item scores that are greater than a more able reference group).

Figures A7 and A8 add solid and dashed lines to the raw and smoothed average item scores, corresponding to the regression-corrected reference and focal average item scores. One important feature in these figures is that the smoothing is not only to the frequencies and average item scores, but also to the regression correction itself. The smoothed regression correction in Figure A8 is generally in line with how the regression correction was designed to work. For most scores, it is decreasing the more-able reference groups' average item scores and increasing the less-able focal groups' average item scores. The exception is at valid subtest scores greater than 32 where the focal series begins to decrease. At these scores the regression correction reduces the focal groups' average item scores.

Consider the raw regression correction in Figure A7. All of the erraticness of the focal examinees' average item scores is also affecting the regression correction. The focal regression correction is extreme at valid subtest scores of 9, 11, 18, 22, 23, 26, 28, 29, and 31. At scores of 9 and 29, the focal average item scores are decreased by the regression correction(because of decreases in average item scores at subtest scores from 8 to 10 and 28 to 30 (note that there were actual data at a subtest score of 8, which was excluded by using the original SIBTEST rules based on the reference item score having a variance of zero). At scores of 26, 28, 29, and 31 where focals' average item scores were abnormally high because they are based on abnormally small raw frequencies for the focals who got Item 22 incorrect (Figure A4), the regression correction increases the average item scores so their difference with the references' regression-corrected average item scores are greater, more positive, and in the opposite direction of what a less-able focal groups' average item scores would be expected to be. The net result is that the

regression correction on the raw average item scores is magnifying the focal-advantaging and reference-advantaging DIF in the raw data and doing so in ways that are inconsistent with the assumptions and intent of the regression correction.

Figure A9 plots the smoothed and raw pieces of the SIBTEST statistic $\hat{B}_{UNI,k}$ that are summed over *k* to form the final $\hat{B}_{UNI}$ statistics. The $\hat{B}_{UNI,k}$ values in Figure 12 reflect the weights based on the total reference and focal data at each *k*. From Figure 12 the explanation for why the raw $\hat{B}_{UNI}$ was much smaller than the smoothed $\hat{B}_{UNI}$ is mostly in the differences in how the raw and smoothed item scores and regression corrections worked in the middle of the data though, to a smaller extent, the smoothing's extrapolation at the high end of the score range is also contributing to a larger smoothed $\hat{B}_{UNI}$. The raw $\hat{B}_{UNI,k}$ values at $k = 23$, 26 and 28 were negative and relatively large, suggesting that the regression-corrected and sample size weighted focal item scores were larger than the corresponding reference item scores. The negative $\hat{B}_{UNI,k}$ values corresponded to places in *k* where the focal *incorrect* data were abnormally small (though not zero) and where the raw regression correction was working erratically and inconsistently. The summing of the $\hat{B}_{UNI,k}$'s resulted in a raw population $\hat{B}_{UNI}$ value that was considerably smaller than the smoothed $\hat{B}_{UNI}$ value.

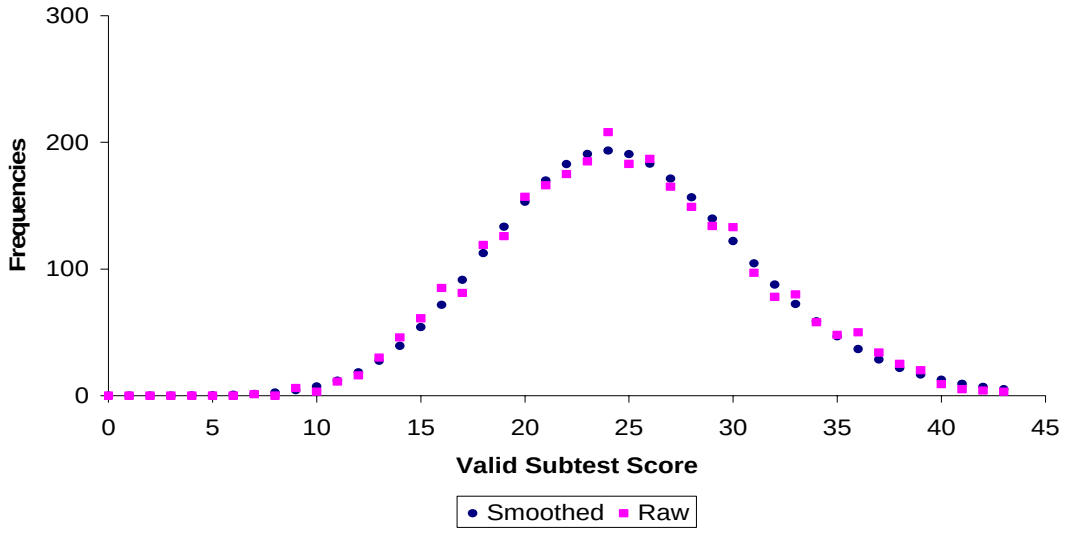

Figure A1. **Smoothed and raw frequencies for the reference examinees who got Item 22 correct.**

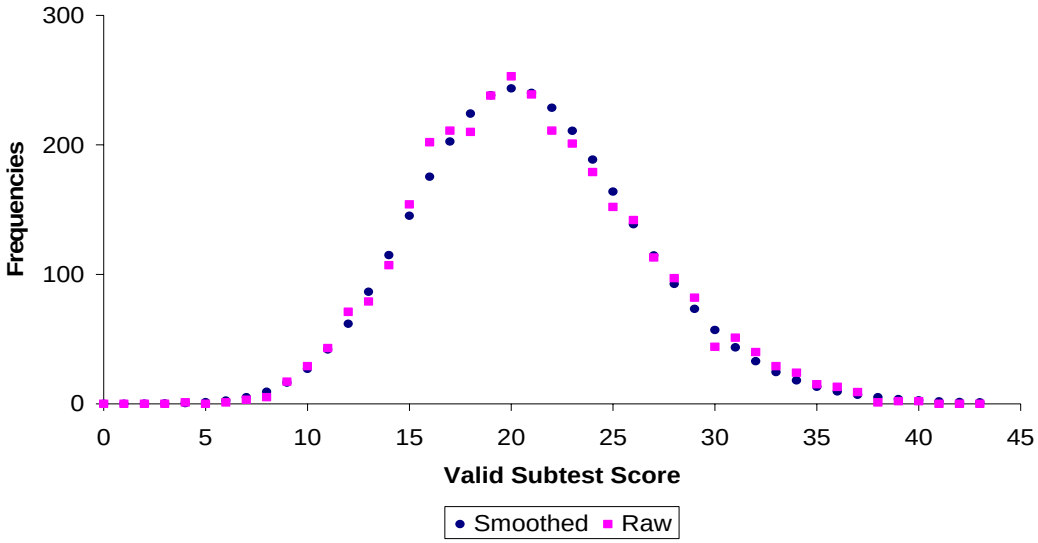

Figure A2. **Smoothed and raw frequencies for the reference examinees who got Item 22 incorrect.**

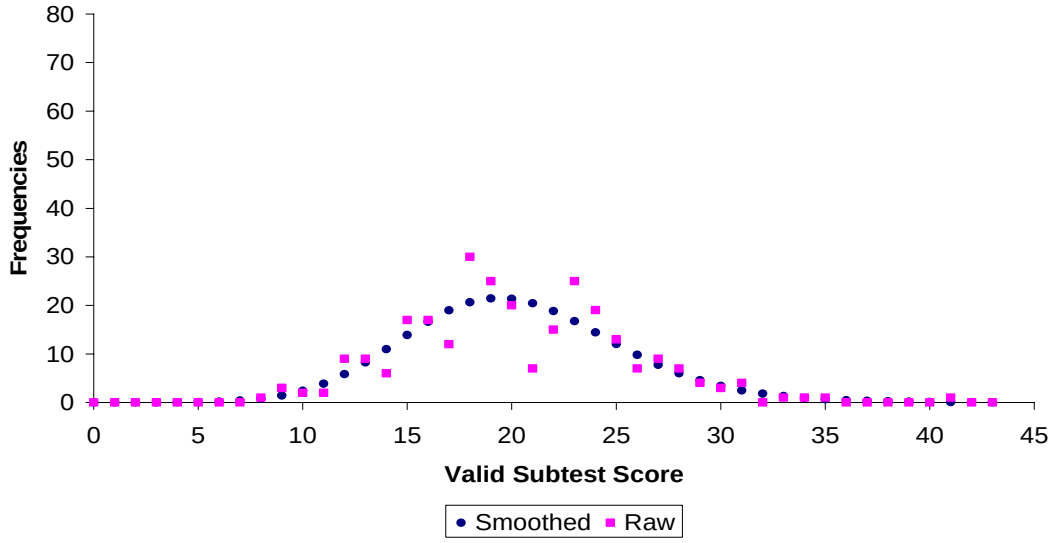

Figure A3. **Smoothed and raw frequencies for the focal examinees who got Item 22 correct.**

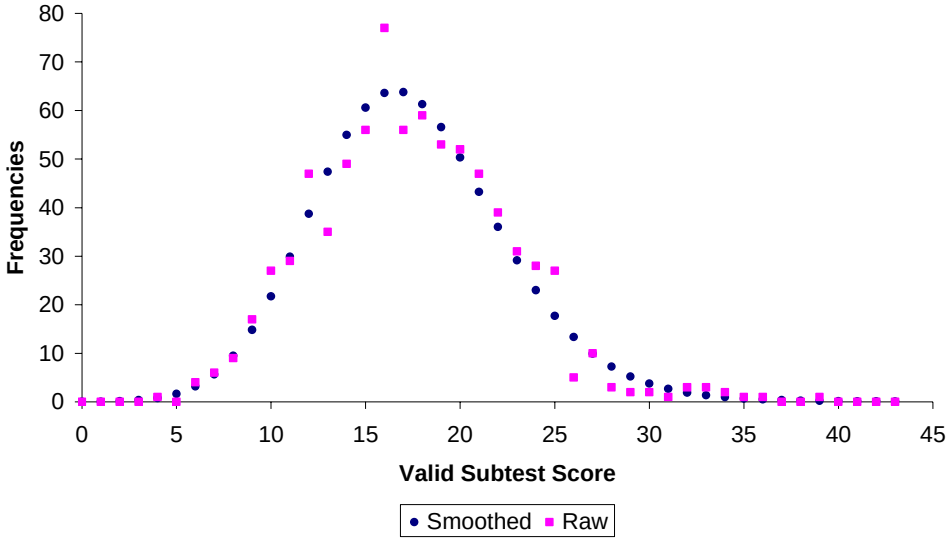

Figure A4. **Smoothed and raw frequencies for the focal examinees who got Item 22 incorrect.**

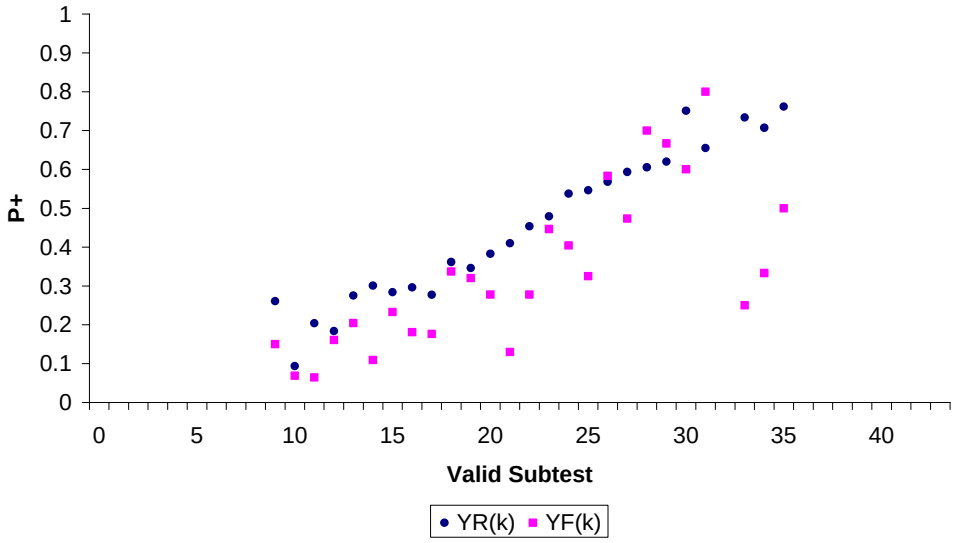

Figure A5. **Raw average item scores.**

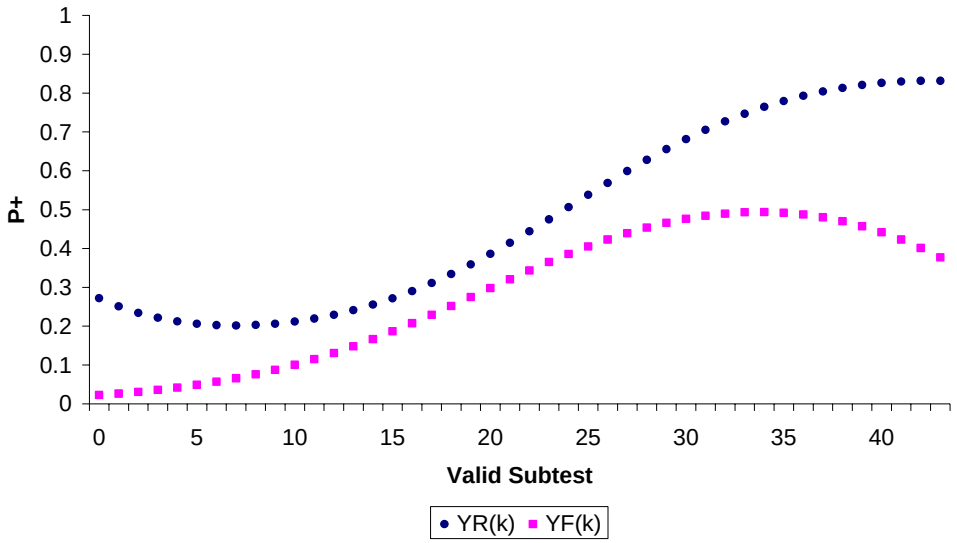

Figure A6. **Smoothed average item scores.**

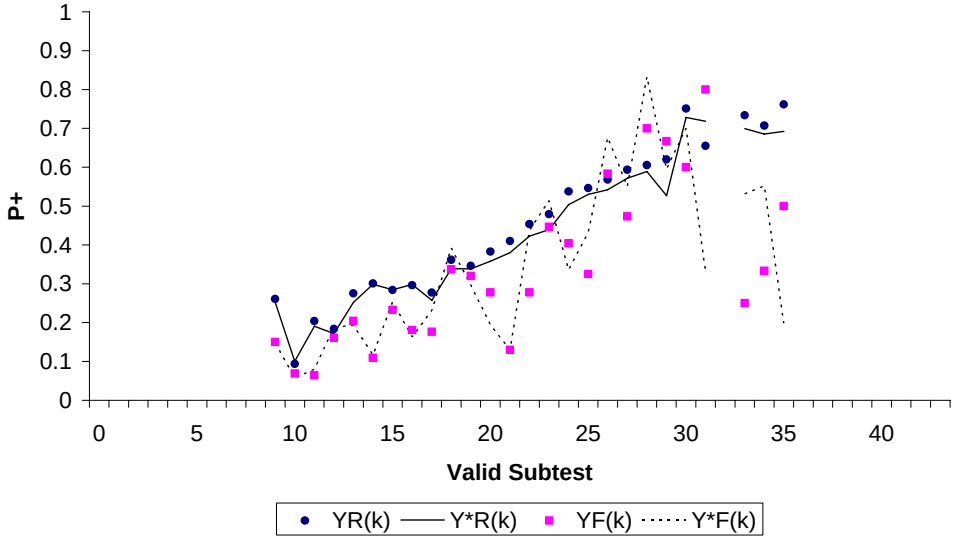

Figure A7. **Raw and regression-corrected item scores.**

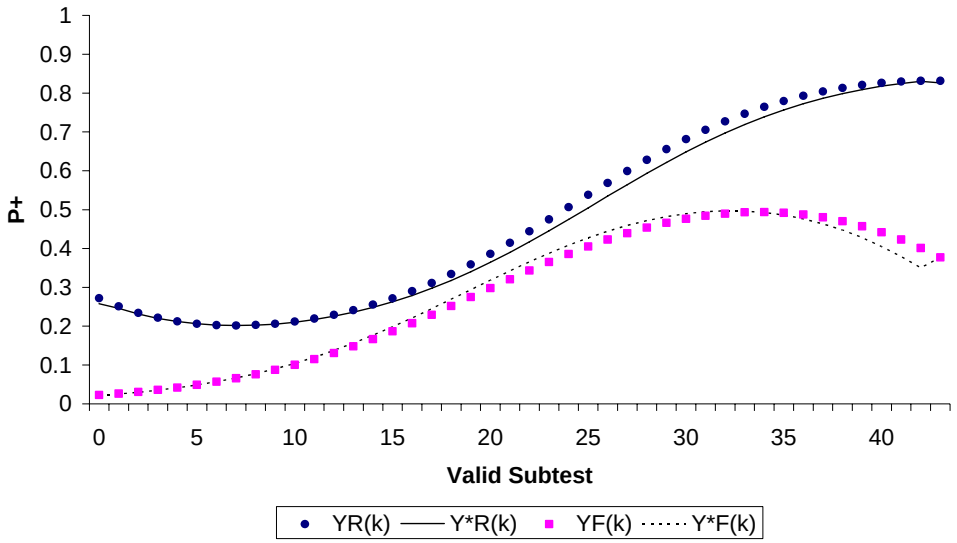

Figure A8. **Smoothed and regression-corrected average item scores.**

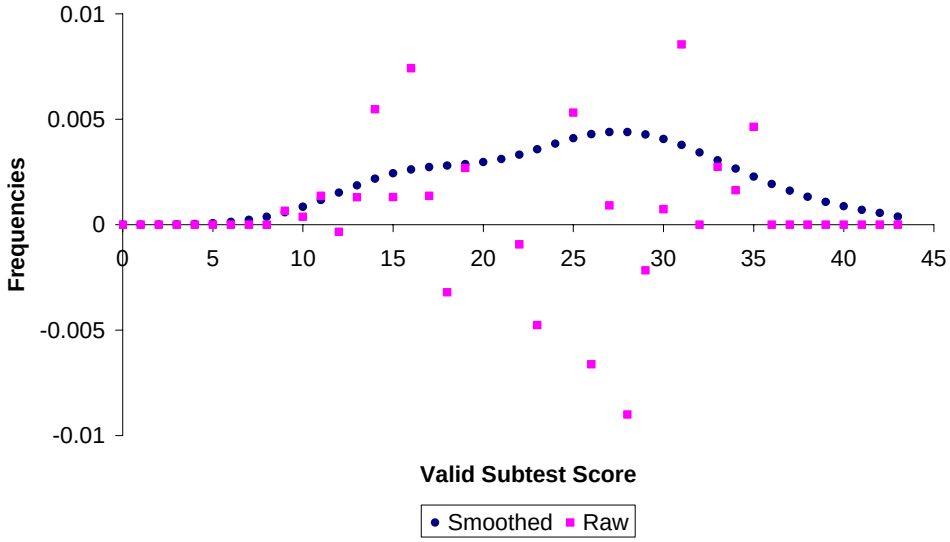

Figure A9. **Smoothed and raw SIBTEST betas (weighted by the total reference and focal data).**